

Feedback Does Not Increase the Capacity of Compound Channels with Additive Noise

Sergey Loyka, Charalambos D. Charalambous

Abstract

A discrete compound channel with memory is considered, where no stationarity, ergodicity or information stability is required, and where the uncertainty set can be arbitrary. When the discrete noise is additive but otherwise arbitrary and there is no cost constraint on the input, it is shown that the causal feedback does not increase the capacity. This extends the earlier result obtained for general single-state channels with full transmitter (Tx) channel state information (CSI) to the compound setting. It is further shown that, for this compound setting and under a mild technical condition on the additive noise, the addition of the full Tx CSI does not increase the capacity either, so that the worst-case and compound channel capacities are the same. This can also be expressed as a saddle-point in the information-theoretic game between the transmitter (who selects the input distribution) and the nature (who selects the channel state), even though the objective function (the inf-information rate) is not convex/concave in the right way. Cases where the Tx CSI does increase the capacity are identified.

Conditions under which the strong converse holds for this channel are studied. The ergodic behaviour of the worst-case noise in otherwise information-unstable channel is shown to be both sufficient and necessary for the strong converse to hold, including feedback and no feedback cases.

I. INTRODUCTION

Many channels, especially wireless ones, are non-ergodic, non-stationary in nature [1] so that the standard tools developed for stationary ergodic channels do not apply and new methods are needed for such channels. A powerful method to deal with general channels, for which

The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Barcelona, Spain, July 2016.

S. Loyka is with the School of Electrical Engineering and Computer Science, University of Ottawa, Ontario, Canada, e-mail: sergey.loyka@ieee.org

C.D. Charalambous is with the ECE Department, University of Cyprus, Nicosia, Cyprus, e-mail: chadcha@ucy.ac.cy

stationarity, ergodicity or information stability are not required, is the information density (spectrum) approach [2][3]. In this method, the key quantity is the inf-information rate rather than the mutual information since the latter does not have operational meaning for information-unstable channels.

In real systems, channel state information (CSI) may be inaccurate or limited due to a number of reasons such as the limitations of channel estimation or feedback link [1]. The concept of compound channel is one way to address this issue whereby a codebook is designed to work for any channel in the uncertainty set, without any knowledge of what channel state is currently in effect [4]. A number of results have been obtained for the capacity of compound channels, see [4] for a detailed review. While most of the studies do not consider feedback, the compound capacity of a class of finite-state memoryless (and hence information-stable) channels with deterministic feedback was established in [5].

While most of the known results require some form of information stability for any channel in the uncertainty set, a general formula for compound channel capacity has been established in [6][8] where no stationarity, ergodicity or information stability is required, and the uncertainty set can be arbitrary. The key quantity in this setting is the compound inf-information rate, which is an extension of the inf-information rate of [2][3] to the compound setting.

In this paper, we extend the study in [6][8] and consider a general compound channel with memory and additive noise (no information stability is required so that the channel can be non-stationary, non-ergodic; the uncertainty set can be arbitrary), where all alphabets are discrete, there is no cost constraint and a noiseless, causal feedback link is present, where all past channel outputs are fed back to the transmitter. We consider a scenario where no CSI is available at the transmitter but full CSI is available to the receiver. Under this setting, we demonstrate that the feedback does not increase the compound channel capacity¹. This extends the earlier result in [7] established for single-state fully-known channels (full CSI available at both ends) to the compound setting. Since noisy feedback cannot outperform noiseless one, this also holds for the former case.

Under a mild technical condition on the additive noise, we further show that the availability

¹In this paper, we consider the classical compound setting [1][4][5] where a fixed-rate code is designed to operate on any channel in the uncertainty set and its decoding regions are allowed to depend on the state (but not the encoding process); variable-rate coding, while being interesting, is beyond the paper's scope.

of the full Tx CSI does not increase the capacity either: the worst-case and compound channel capacities are the same. This fact is remarkable since achieving the worst-case capacity allows for the codebooks to depend on the channel state while the compound channel capacity requires the codebooks to be independent of the channel state (and hence no feedback to the Tx is needed). This can also be expressed as the existence of a saddle point in the information-theoretic game between the transmitter (who selects the input) and the nature (who selects the channel state): neither player can deviate from the optimal strategy without incurring a penalty. This result is rather surprising since the underlying objective function (the inf-information rate) is *not* convex/concave in the right way and the uncertainty set can be non-convex as well (e.g. discrete) so that the celebrated von Neumann's mini-max Theorem [15] or its extensions [16] cannot be used to establish the existence of a saddle-point. This shows that neither convexity of the feasible set nor of the objective function are necessary for a saddlepoint to exist in this information-theoretic game. This saddlepoint result extends the earlier results established for stationary and ergodic (and hence information-stable) channels, e.g. in [17]-[22], where mutual information is a proper metric, to the realm of information-unstable scenarios, where the inf-information rate has to be used as a metric since the mutual information does not have operational meaning anymore.

Next, we consider some cases when the Tx CSI does increase the capacity. This turns out to be somewhat surprising since, in all such cases, the optimal input distribution is uniform, regardless of the channel state (a common wisdom suggests that the Tx CSI increases the capacity via proper selection of the input distribution tailored to the channel state; our results indicate that this does not have to be the case). Examples are provided to facilitate understanding and insights.

Finally, conditions under which the strong converse holds for this channel are studied. The ergodic behaviour of the worst-case noise in otherwise information-unstable channel is shown to be both sufficient and necessary for the strong converse to hold, including feedback and no feedback cases. Examples are given to illustrate scenarios when the strong converse holds and when it does not.

The rest of the paper is organized as follows. Section II introduces the channel model and notations. Section III discusses the capacity of general (information-unstable) compound channels without feedback. The impact of feedback is included in Section IV. The impact of the channel state information at the transmitter and the existence of a saddlepoint are studied in Section V.

Examples are given in Section VI. Sufficient and necessary conditions for the strong converse to hold are given in Section VII.

II. CHANNEL MODEL

Let us consider the following discrete-time model of a compound discrete channel with additive noise:

$$y^n = g_s^n(x^n) + \xi_s^n \quad (1)$$

where x^n, y^n, ξ_s^n are the input, output and noise sequences of length n , $x^n = \{x_1, \dots, x_n\}$ and likewise for y^n and ξ_s^n ; the functions $g_s^n(x^n) = \{g_{s1}(x_1), g_{s2}(x_2), \dots, g_{sn}(x_n)\}$ model the channel's impulse response and are required to induce one-to-one mapping $x^n \leftrightarrow z^n = g_s^n(x^n)$. All alphabets as well as operations are M -ary, $s \in \mathcal{S}$ denotes the channel (noise) state, and \mathcal{S} is the (arbitrary) channel uncertainty set. The compound sequence $\xi_s^n = \{\xi_{1s}, \dots, \xi_{ns}\}$ represents arbitrary additive noise, e.g. non-ergodic, non-stationary in general, independent of the channel input when used without feedback.

Note that the channel is not memoryless, it may include inter-symbol interference (ISI) via $g_s^n(\cdot)$, e.g.

$$z_k = g_{sk}(x^k) = \sum_{i=0}^{l_s} x_{k-i} \quad (2)$$

where l_s is the depth of the ISI and where we set $x_i = 0$ for $i < 0$. The noise is also allowed to have arbitrary memory.

The channel is not required to be information stable (in the sense of Dobrushin [12] or Pinsker [13]). We assume that s is known to the receiver but not the transmitter, who knows the (arbitrary) uncertainty set \mathcal{S} . This is motivated by the fact that channel estimation is done at the receiver; M may be small, e.g. binary alphabets, while the cardinality of \mathcal{S} can be very large (in fact, \mathcal{S} can be a continuous set) so it is not feasible in practice to feed s back to the transmitter via e.g. a binary feedback channel.

The channel has noiseless feedback with 1-symbol delay (which can also be extended to noisy feedback - see Remark 3), so that the transmitted symbol x_k at time $k = 1..n$ is selected as $x_k^{(n)} = f_k^{(n)}(w^n y^{k-1})^2$, where n is the blocklength, w^n denotes the message to be transmitted

²our result will also hold for a more general feedback of the form $u_k = \beta_k(y^k)$, where $\{\beta_k\}$ are arbitrary feedback functions, see Remark 4.

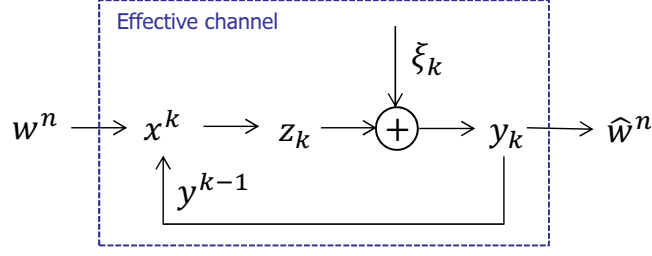


Fig. 1. A general channel with additive noise and causal feedback $\{w^n y^{k-1}\} \rightarrow x^k \rightarrow y_k$ and the effective channel $w^n \rightarrow y^n$ (dashed box), where $z_k = g_{sk}(x^k)$, $x^k = f^k(w^n y^{k-1})$.

via n -symbol block, $x^n = \{x_1^{(n)} \dots x_n^{(n)}\}$ and likewise for y^n and ξ_s^n ; $f_k^{(n)}$ denotes the encoding function at time k , which depends on the selected message w^n and past channel outputs y^{k-1} (due to the feedback); $f^n = \{f_1^{(n)} \dots f_n^{(n)}\}$. This induces the input distribution of the form:

$$p(x^n || y^{n-1}) = \prod_{k=1}^n p(x_k | x^{k-1} y^{k-1}) \quad (3)$$

where $||$ denotes causal conditioning [10]. No cost constraint is imposed on the input.

Notations: To simplify notations, we use $p(x|y)$ to denote conditional distribution $p_{x|y}(x|y)$ when this causes no confusion (and likewise for joint and marginal distributions) and shortcut $x_k^{(n)}$ as x_k with understanding that all sequences and distributions depend on blocklength n and may be different for different blocklengths. Capitals (X) denote random variables while lower-case letters (x) denote their realizations or arguments of functions; $\mathbf{X} = \{X^n\}_{n=1}^\infty$.

III. CAPACITY WITHOUT FEEDBACK

First, we briefly review the relevant capacity result in [6][8], which apply to general compound channels $p_s(y^n|x^n)$, not only those in (1); channels can be information-unstable, e.g. non-stationary, non-ergodic, but without feedback, i.e. $x^k = f^k(w^n)$ (the input depends only on the message and the past inputs, not the outputs). The compound channel capacity is defined operationally as the maximum achievable rate for which the error probability can be made arbitrary small and uniformly so over the whole set of channels and where the codewords are independent of channel state (see e.g. [4][8] for details).

Theorem 1 ([6][8]). Consider a general compound channel where the channel state $s \in \mathcal{S}$ is known to the receiver but not the transmitter and is independent of the channel input; the transmitter knows the (arbitrary) uncertainty set \mathcal{S} . Its compound channel capacity (without feedback) is given by

$$C_{NFB} = \sup_{\mathbf{X}} \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \quad (4)$$

where the supremum is over all sequences of finite-dimensional input distributions and $\underline{\underline{I}}(\mathbf{X}; \mathbf{Y})$ is the compound inf-information rate,

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) = \sup_R \left\{ R : \lim_{n \rightarrow \infty} \sup_{s \in \mathcal{S}} \Pr \{Z_{ns} \leq R\} = 0 \right\} \quad (5)$$

where $Z_{ns} = n^{-1}i(X^n; Y^n|s)$ is the normalized information density under channel state s . \square

This theorem was proved in [6][8] using the Verdu-Han and Feinstein Lemmas properly extended to the compound channel setting.

For future use, we need the following formal definitions, which extend the respective inf and sup operators introduced for regular (single-state) sequences [2][3] (see (75)) to the compound setting.

Definition 1. Let $\{X_{sn}\}_{n=1}^{\infty}$ be an arbitrary compound random sequence where s is a state (i.e. a random sequence indexed by the state s). The compound infimum $\underline{\underline{\{\cdot\}}}$ and supremum $\overline{\overline{\{\cdot\}}}$ operators are defined as follows:

$$\underline{\underline{\mathbf{X}}} = \underline{\underline{\{X_{sn}\}}} = \sup \left\{ x : \lim_{n \rightarrow \infty} \sup_s \Pr \{X_{sn} \leq x\} = 0 \right\} \quad (6)$$

$$\overline{\overline{\mathbf{X}}} = \overline{\overline{\{X_{sn}\}}} = \inf \left\{ x : \lim_{n \rightarrow \infty} \sup_s \Pr \{X_{sn} \geq x\} = 0 \right\} \quad (7)$$

Roughly, $\underline{\underline{\mathbf{X}}}$ and $\overline{\overline{\mathbf{X}}}$ represent the largest lower and least upper bounds to the asymptotic support set of X_{sn} over the whole state set (note \sup_s in the definitions).

The following definitions extend the respective information-theoretic quantities in [2][3] to the compound setting.

Definition 2. Let $\mathbf{X} = \{X_s^n\}_{n=1}^{\infty}$ and $\mathbf{Y} = \{Y_s^n\}_{n=1}^{\infty}$ be two compound random sequences with distributions p_{sx^n} and p_{sy^n} where s is a state. The compound inf-divergence rate is defined as

$$\underline{\underline{D}}(\mathbf{X}; \mathbf{Y}) = \underline{\underline{\{d_{sn}(X_s^n || Y_s^n)\}}} \quad (8)$$

where $d_{sn}(x^n||y^n) = \frac{1}{n} \log \frac{p_{sx^n}(x^n)}{p_{sy^n}(x^n)}$ is the divergence density rate. The compound inf and sup-entropy rates $\underline{\underline{H}}(\mathbf{X})$ and $\overline{\overline{H}}(\mathbf{X})$ are defined as

$$\underline{\underline{H}}(\mathbf{X}) = \underline{\underline{\{h_{sn}(X_s^n)\}}}, \quad \overline{\overline{H}}(\mathbf{X}) = \overline{\overline{\{h_{sn}(X_s^n)\}}} \quad (9)$$

where $h_{sn}(x^n) = -n^{-1} \log p_s(x^n)$ is the entropy density rate. The compound conditional inf-entropy rate $\underline{\underline{H}}(\mathbf{Y}|\mathbf{X})$ and sup-entropy rate $\overline{\overline{H}}(\mathbf{Y}|\mathbf{X})$ are defined analogously via the conditional entropy density rate $h_{sn}(y^n|x^n) = -n^{-1} \log p_s(y^n|x^n)$ (with respect to the joint distribution $p_s(x^n, y^n)$), and $\overline{\overline{I}}(\mathbf{X}; \mathbf{Y})$ is similarly defined.

The proposition below gives the properties of the compound inf-information rate $\underline{\underline{I}}(\mathbf{X}; \mathbf{Y})$ and other relevant quantities [6][8], which will be instrumental below.

Proposition 1. *Let \mathbf{X} and \mathbf{Y} be (arbitrary) compound random sequences. The following holds:*

$$\underline{\underline{D}}(\mathbf{X}||\mathbf{Y}) \geq 0 \quad (10)$$

$$\overline{\overline{I}}(\mathbf{X}; \mathbf{Y}) \geq \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \geq 0 \quad (11)$$

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) = \underline{\underline{I}}(\mathbf{Y}; \mathbf{X}) \quad (12)$$

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \leq \overline{\overline{H}}(\mathbf{Y}) - \overline{\overline{H}}(\mathbf{Y}|\mathbf{X}) \quad (13)$$

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \leq \underline{\underline{H}}(\mathbf{Y}) - \underline{\underline{H}}(\mathbf{Y}|\mathbf{X}) \quad (14)$$

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \geq \underline{\underline{H}}(\mathbf{Y}) - \overline{\overline{H}}(\mathbf{Y}|\mathbf{X}) \quad (15)$$

$$\overline{\overline{H}}(\mathbf{Y}) \geq \overline{\overline{H}}(\mathbf{Y}|\mathbf{X}) \quad (16)$$

$$\overline{\overline{H}}(\mathbf{Y}) \geq \underline{\underline{H}}(\mathbf{Y}) \geq \underline{\underline{H}}(\mathbf{Y}|\mathbf{X}) \quad (17)$$

If the alphabets are discrete, then

$$0 \leq \underline{\underline{H}}(\mathbf{X}|\mathbf{Y}) \leq \underline{\underline{H}}(\mathbf{X}) \leq \overline{\overline{H}}(\mathbf{X}) \leq \log N_x \quad (18)$$

$$\begin{aligned} 0 \leq \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) &\leq \min\{\underline{\underline{H}}(\mathbf{X}), \underline{\underline{H}}(\mathbf{Y})\} \\ &\leq \min\{\log N_x, \log N_y\} \end{aligned} \quad (19)$$

$$\begin{aligned} \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) &= \min\{\underline{\underline{H}}(\mathbf{X}), \underline{\underline{H}}(\mathbf{Y})\} \\ &\text{if } \min\{\overline{\overline{H}}(\mathbf{Y}|\mathbf{X}), \overline{\overline{H}}(\mathbf{X}|\mathbf{Y})\} = 0 \end{aligned} \quad (20)$$

$$\begin{aligned} 0 \leq \overline{\overline{I}}(\mathbf{X}; \mathbf{Y}) &\leq \min\{\overline{\overline{H}}(\mathbf{X}), \overline{\overline{H}}(\mathbf{Y})\} \\ &\leq \min\{\log N_x, \log N_y\} \end{aligned} \quad (21)$$

where the last inequalities in (18)-(21) hold if the alphabets are of finite cardinality N_x, N_y .

Note that many of these properties mimic the respective properties of mutual information and entropy, e.g. "conditioning cannot increase the entropy" and "mutual information is non-negative, symmetric and bounded by the entropy of the alphabet".

IV. CAPACITY WITH FEEDBACK

In this section, we consider a discrete compound channel with feedback and general additive noise. Instead of dealing with the feedback channel $\{w^n y^{k-1}\} \rightarrow x^k \rightarrow y_k, k = 1 \dots n$, directly, one can consider an effective channel $w^n \rightarrow y^n$ without feedback, see Fig. 1. Applying Theorem 1 to the effective channel, the capacity with the feedback can be expressed as³

$$C_{FB} = \sup_{\mathbf{W}, \mathbf{F}} \underline{\underline{I}}(\mathbf{W}; \mathbf{Y}) \quad (22)$$

where $\mathbf{Y} = \{Y^n\}_{n=1}^\infty$ and likewise for \mathbf{W} , $\underline{\underline{I}}(\mathbf{W}; \mathbf{Y})$ is the compound inf-information rate:

$$\underline{\underline{I}}(\mathbf{W}; \mathbf{Y}) = \underline{\underline{\{n^{-1}i(W^n; Y^n|_s)\}}} \quad (23)$$

where $i(W^n; Y^n|_s)$ is the information density:

$$i(w^n; y^n|_s) = \log \frac{p_s(y^n|w^n)}{p_s(y^n)} \quad (24)$$

³see also [11] for a formulation based on the directed information for the case of full CSI and a proof of equivalence of these two formulations in the latter case.

The maximization in (22) is over all possible encoding functions $\mathbf{F} = \{f^n\}_{n=1}^\infty$ and all possible message distributions. Unfortunately, this maximization is difficult to perform in general. Therefore, we proceed in a different way. Let

$$\overline{\overline{H}}(\Xi) = \overline{\overline{\{n^{-1}h(\Xi_s^n|s)\}}} \quad (25)$$

be the compound sup-entropy rate of the compound noise $\Xi = \{\Xi_s^n\}_{n=1}^\infty$, $\Xi_s^n = \{\Xi_{1s} \dots \Xi_{ns}\}$, $h(\xi^n|s) = -\log p_s(\xi^n)$. The following is the main result of the paper.

Theorem 2. *The capacity of the compound discrete channel with (arbitrary) additive noise in (1) and the full Rx CSI is not increased by the causal feedback:*

$$C_{FB} = C_{NFB} = \sup_{\mathbf{X}} \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) = \log M - \overline{\overline{H}}(\Xi) \quad (26)$$

where C_{NFB} is the capacity without feedback, and $\sup_{\mathbf{X}}$ is over all sequences of input distributions.

Proof: Let us consider the no-feedback case first. The 2nd equality follow from Theorem 1. The following Lemma is needed to prove the last equality.

Lemma 1. *Let $z_k = g_{sk}(x^k)$, $k = 1 \dots n$, and the mapping $x^n \rightarrow z^n$ is one-to-one. If $p(x^n) = 1/M^n$, then $p(z^n) = 1/M^n$, i.e. equiprobable X^n generates equiprobable Z^n .*

Now, since the mapping $x^n \rightarrow z^n$ is invertible, $\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) = \underline{\underline{I}}(\mathbf{Z}; \mathbf{Y})$. It follows from (13) that

$$\underline{\underline{I}}(\mathbf{Z}; \mathbf{Y}) \leq \overline{\overline{H}}(\mathbf{Y}) - \overline{\overline{H}}(\mathbf{Y}|\mathbf{Z}) \quad (27)$$

Using this inequality, one obtains:

$$\underline{\underline{I}}(\mathbf{Z}; \mathbf{Y}) \leq \log M - \overline{\overline{H}}(\mathbf{Y}|\mathbf{Z}) \quad (28)$$

$$= \log M - \overline{\overline{H}}(\Xi) \quad (29)$$

where 1st inequality is due to M -ary alphabets, so that $\overline{\overline{H}}(\mathbf{Y}) \leq \log M$ (see (18)), and the equality is due to $\overline{\overline{H}}(\mathbf{Y}|\mathbf{Z}) = \overline{\overline{H}}(\mathbf{Z} + \Xi|\mathbf{Z}) = \overline{\overline{H}}(\Xi)$, since the noise is additive and independent of the input (recall that we consider the no feedback case). Finally,

$$\underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) \leq \log M - \overline{\overline{H}}(\Xi) \quad (30)$$

and the equality is achieved by equiprobable input due to Lemma 1, under which the output is also equiprobable. This proves the last equality in (26).

To prove 1st equality, $C_{FB} = C_{NFB}$, observe that feedback cannot decrease the capacity,

$$\log M - \overline{\overline{H}}(\Xi) = C_{NFB} \leq C_{FB} \quad (31)$$

To prove the converse,

$$C_{FB} \leq \log M - \overline{\overline{H}}(\Xi) \quad (32)$$

use (22) to conclude

$$C_{FB} = \sup_{\mathbf{W}, \mathbf{F}} \underline{\underline{I}}(\mathbf{W}; \mathbf{Y}) \quad (33)$$

$$\leq \sup_{\mathbf{W}, \mathbf{F}} [\overline{\overline{H}}(\mathbf{Y}) - \overline{\overline{H}}(\mathbf{Y}|\mathbf{W})] \quad (34)$$

$$\leq \log M - \inf_{\mathbf{W}, \mathbf{F}} \overline{\overline{H}}(\mathbf{Y}|\mathbf{W}) \quad (35)$$

where 1st inequality is due to $\underline{\underline{I}}(\mathbf{W}; \mathbf{Y}) \leq \overline{\overline{H}}(\mathbf{Y}) - \overline{\overline{H}}(\mathbf{Y}|\mathbf{W})$ and 2nd inequality is due to $\overline{\overline{H}}(\mathbf{Y}) \leq \log M$ (since the alphabet is M -ary).

To evaluate $\overline{\overline{H}}(\mathbf{Y}|\mathbf{W})$, note that

$$p_s(y^n|w^n) = \prod_{k=1}^n p_s(y_k|y^{k-1}w^n) \quad (36)$$

and

$$p_s(y_k|y^{k-1}w^n) = p_s(y_k|y^{k-1}x^k w^n) \quad (37)$$

$$= p_s(y_k|y^{k-1}x^k \xi^{k-1} w^n) \quad (38)$$

$$= p_{sy}(g_{sk}(x^k) + \xi_k | x^k \xi^{k-1} w^n) \quad (39)$$

$$= p_{s\xi}(\xi_k | \xi^{k-1} w^n) \quad (40)$$

$$= p_{s\xi}(\xi_k | \xi^{k-1}) \quad (41)$$

where $\xi_k = y_k - g_{sk}(x^k)$, $x^k = f^k(w^n y^{k-1})$, $\xi^n = \{\xi_k\}_{k=1}^n$. 1st equality is due to $x^k = f^k(w^n y^{k-1})$; 2nd and 3rd equalities are due to the channel model $y_k = z_k + \xi_k$; 4th equality is due to $x^k = \check{f}^k(w^n \xi^{k-1})$, where \check{f}^k is a function which depends on encoding functions f^k and

the channel impulse response functions g_s^k ; last equality is due to independence of noise and message. Thus,

$$p_{sy}(y^n|w^n) = p_{s\xi}(\xi^n) \quad (42)$$

and therefore

$$\overline{\overline{H}}(\mathbf{Y}|\mathbf{W}) = \overline{\overline{H}}(\Xi) \quad (43)$$

Combining this with (35), one obtains (32) and hence the desired result follows.

Equality in (32) is achieved by the uniform input distribution $p(x^n) = 1/M^n$, which is also i.i.d. and equiprobable: $p(x^n) = \prod_{i=1}^n p(x_i)$, $p(x_i) = 1/M$ (this can be shown by induction), under which the output is also uniform. ■

Remark 1. *Note that in both feedback and no-feedback systems, the optimizing input is uniform and hence i.i.d. equiprobable and independent of the feedback. Under this input, the output is also uniform under any noise, which explains why feedback is not helpful in this setting.*

Remark 2. *Setting $l_s = 0$ in (2), one obtains a channel without intersymbol interference. When, in addition, the uncertainty set \mathcal{S} is singleton (single-state channel with no uncertainty), Theorem 2 above reduces to the corresponding result in [7] obtained for fully known (no uncertainty) channels.*

Remark 3. *Since noisy feedback cannot perform better than noiseless, this result also implies that noisy feedback cannot increase the compound capacity in this setting either.*

Remark 4. *One may consider a more general feedback of the form $u_k = \beta_k(y^k)$, where $\{\beta_k\}$ are arbitrary (possibly random) feedback functions (which account for e.g. quantization of feedback signals and noise in the feedback channel), and the corresponding encoding of the form $x_k = f_k(w^n u^{k-1})$. Since the capacity with this form of feedback cannot exceed the capacity with the full feedback of y^{k-1} , Theorem 2 still holds for this setting as well.*

V. IMPACT OF THE TX CSI AND SADDLE POINT

Let us consider the case where channel state s is known at the transmitter, so that codewords can be selected as functions of the channel state. In this case, the worst-case channel capacity

C_w is a proper performance metric and it can be expressed as

$$C_w = \inf_s \sup_{\mathbf{W}, \mathbf{F}} \underline{I}(\mathbf{W}; \mathbf{Y}|s) \quad (44)$$

$$= \inf_s (\log M - \overline{H}(\Xi|s)) \quad (45)$$

$$= \inf_s \sup_{\mathbf{X}} \underline{I}(\mathbf{X}; \mathbf{Y}|s) \quad (46)$$

$$= \log M - \sup_s \overline{H}(\Xi|s) \quad (47)$$

$$\geq \log M - \overline{\overline{H}}(\Xi) = C_{FB} \quad (48)$$

where $\underline{I}(\mathbf{X}; \mathbf{Y}|s)$ is the inf-information rate under channel state s [2]:

$$\underline{I}(\mathbf{X}; \mathbf{Y}|s) = \sup_R \left\{ R : \lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} i(W^n; Y^n|s) \leq R \right\} = 0 \right\}, \quad (49)$$

$\overline{H}(\Xi|s)$ is the sup-entropy rate of the noise under state s :

$$\overline{H}(\Xi|s) = \inf_R \left\{ R : \lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} h(\Xi_s^n|s) > R \right\} = 0 \right\} \quad (50)$$

(44) follows from the general formula in [2] and the equivalent channel in Fig. 1; (45) follows from the Theorem in [7]; (48) follows from the Lemma 2 below, so that the impact of Tx CSI can be characterized by

$$\Delta C = C_w - C_{FB} = \overline{\overline{H}}(\Xi) - \sup_s \overline{H}(\Xi|s) \geq 0 \quad (51)$$

Note that, similarly to the compound capacity, C_w is not increased by the feedback either, i.e. (47) is also the no-feedback worst-case channel capacity as indicated by (46) while $C_s = \sup_{\mathbf{X}} \underline{I}(\mathbf{X}; \mathbf{Y}|s)$ is the channel capacity under state s known to both Tx and Rx.

To proceed further, we need the following definition.

Definition 3. The compound noise sequence $\{\Xi_s^n\}_{n=1}^\infty$ is uniform if the convergence in

$$\Pr \left\{ \frac{1}{n} h(\Xi_s^n|s) > \sup_s \overline{H}(\Xi|s) + \delta \right\} \rightarrow 0 \quad (52)$$

as $n \rightarrow \infty$ is uniform in $s \in \mathcal{S}$ for any $\delta > 0$.

Note that, while the convergence to zero in (52) for each $\delta > 0$ and $s \in \mathcal{S}$ is guaranteed from the definition of $\sup_s \overline{H}(\Xi|s)$, this convergence does not have to be uniform in general. In fact, the uniform convergence requirement above is equivalent to

$$\lim_{n \rightarrow \infty} \sup_s \Pr \left\{ \frac{1}{n} h(\Xi_s^n|s) > \sup_s \overline{H}(\Xi|s) + \delta \right\} = 0 \quad (53)$$

for any $\delta > 0$, which is clearly stronger than just point-wise convergence in (52) for each s , which is equivalent to

$$\sup_s \lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} h(\Xi_s^n | s) > \sup_s \overline{H}(\Xi | s) + \delta \right\} = 0 \quad (54)$$

In general, \lim and \sup cannot be swapped; rather

$$\sup_s \lim_{n \rightarrow \infty} \{\cdot\} \leq \lim_{n \rightarrow \infty} \sup_s \{\cdot\} \quad (55)$$

so that (53) implies (54) but the converse is not true in general, i.e. the inequality can be strict.

We are now in a position to establish the following key result.

Lemma 2. *The following inequality holds for the general compound noise sequence:*

$$\overline{\overline{H}}(\Xi) \geq \sup_s \overline{H}(\Xi | s) \quad (56)$$

with equality if and only if the compound noise is uniform.

Proof: See Appendix. ■

Strict inequality in (56) can be demonstrated by examples - see Section VI. Combining Lemma 2 with (51), one obtains the following result.

Theorem 3. *Consider the discrete compound channel with additive noise as in (1) under the full Rx CSI. When the compound noise is uniform, neither the full Tx CSI nor causal noiseless or noisy feedback increase its capacity, i.e.*

$$C_{NFB} = C_{FB} = C_w \quad (57)$$

The last equality states that the worst-case channel capacity (achievable by codebooks tailored to the channel state) is the same as the compound channel capacity (where the codebooks are independent of channel states), which can be equivalently expressed as

$$\inf_s \sup_{\mathbf{X}} \underline{I}(\mathbf{X}; \mathbf{Y} | s) = \sup_s \inf_{\mathbf{X}} \underline{I}(\mathbf{X}; \mathbf{Y} | s) \quad (58)$$

so that, when \inf and \sup are achieved, this is equivalent to the existence of a saddle point [15][16]:

$$\underline{I}(\mathbf{X}; \mathbf{Y} | s^*) \leq \underline{I}(\mathbf{X}^*; \mathbf{Y}^* | s^*) \leq \underline{I}(\mathbf{X}^*; \mathbf{Y}^* | s) \quad (59)$$

where (\mathbf{X}^*, s^*) is the saddle point. This saddle point exists for both feedback and no feedback cases when the compound noise is uniform.

It is remarkable that a saddle point exists even though the uncertainty set is allowed to be non-convex and the objective function $f(s) \triangleq \underline{I}(\mathbf{X}; \mathbf{Y}|s)$ is not required to be convex either (e.g. when s is discrete, $f(s)$ is not convex; it can also be non-convex even when the uncertainty set is convex), so that von Neumann's mini-max Theorem [15] or its extensions [16] can not be used to establish the existence of a saddle point.

The saddle point above extends the information-theoretic saddle-point results established earlier in e.g. [17]-[22] for stationary and ergodic (and hence information-stable) channels, for which mutual information is a proper metric, to the realm of information-unstable scenarios, where mutual information has no operational meaning and the inf-information rate has to be used instead. Furthermore, it demonstrates that neither convexity of the feasible set nor of the objective function are necessary for the saddlepoint to exist. It also has the standard game-theoretic interpretation: neither the nature (who controls state s) nor the transmitter (who controls the input distribution) can deviate from the optimal strategy without incurring a penalty.

VI. EXAMPLES

In this section, we consider some illustrative examples. Among other things, they identify the scenarios when the Tx CSI increases the capacity and when it does not.

A. Example 1

Let the compound noise be of the form

$$\xi_s^n = \{w_1, \dots, w_s, 0..0\} \quad (60)$$

where W_i are i.i.d. equiprobable so that $p(w^s) = 1/M^s$, and $s \in \{1, 2, \dots\}$. This can model block interference/noise of length s . Note that the noise process $\{\xi_s^n\}$ is not stationary. Using (50), one obtains, after some manipulations, $\overline{H}(\Xi|s) = 0 \ \forall s$ (this is due to the fact that, under fixed s , the "noisy" part in (60) is asymptotically negligible) so that $\sup_s \overline{H}(\Xi|s) = 0$ and hence

$$C_w = \log M \quad (61)$$

Yet, using (50), it follows, after some manipulations, that $\overline{H}(\Xi) = \log M > \sup_s \overline{H}(\Xi|s) = 0$. Hence, the noise is not uniform and

$$C_{FB} = 0 \quad (62)$$

so that the advantage of the Tx CSI is significant: $\Delta C = \log M$, i.e. the maximum possible value for M -ary alphabet. The reason for this is that the compound noise in (60) is not uniform, the worst-case noise (corresponding to \sup_s in (6)) is i.i.d. equiprobable for any given n and hence the compound channel is useless, even under noiseless causal feedback. The presence of the Tx CSI changes the situation dramatically: one can now design a codebook for any given state s and make the error probability arbitrary low by using sufficiently large blocklength $n \gg s$. This conclusion also holds for *any* distribution of w_s^n , not only i.i.d. equiprobable, since $\sup_s \overline{H}(\Xi|s) = 0$ regardless.

The situation also changes dramatically if one imposes the boundedness constraint on the uncertainty set: $s \leq S < \infty$. In this case, $\overline{H}(\Xi) = \sup_s \overline{H}(\Xi|s) = 0$, i.e. the noise becomes uniform, and hence $C_{FB} = C_w = \log M$. One may wonder as to what are the practical implications of these dramatic changes. In our view, the first case of unbounded s corresponds to a scenario where the interference is more powerful than the codebook, i.e. for any given n , does not matter how large, one can always find powerful enough interference with $s = n$ thus rendering the channel useless. The second case of bounded s prevents this thus allowing for the codeword length n to be much larger than S and hence represents a scenario where the codebook is more powerful than any possible interference. In the same way, one can interpret the impact of the Tx CSI: giving s to the Tx allows one to chose $n \gg s$ and hence make the impact of interference negligible, which is not possible otherwise.

B. Example 2

Let us now set the compound noise as

$$\xi_s^n = \{w_1, \dots, w_s, z_{s+1}, \dots, z_n\} \quad (63)$$

with binary alphabet and $W_i \sim \text{Ber}(p_1)$, $Z_i \sim \text{Ber}(p_2)$, i.e. Bernoulli random variables, all independent of each other and $0 \leq p_2 < p_1 \leq 1/2$, and $s \in \{1, 2, \dots\}$. This can model a scenario where there is noise (2nd part) in addition to interference (1st part).

One obtains, after some manipulations,

$$\sup_s \overline{H}(\Xi|s) = h(p_2) < h(p_1) = \overline{\overline{H}}(\Xi) \quad (64)$$

where $h(p)$ is the binary entropy function, and hence

$$\Delta C = C_w - C_{FB} = h(p_1) - h(p_2) > 0 \quad (65)$$

so that the noise is not uniform and the Tx CSI does bring in advantage. Bounding the uncertainty set $s \leq S < \infty$ has no impact on $\overline{H}(\Xi|s)$ but makes $\overline{\overline{H}}(\Xi) = h(p_2)$ and hence the advantage of the Tx CSI disappear: $\Delta C = 0$. The noise becomes uniform in this case. As in Example 1, the distribution of w^s does not affect $\overline{H}(\Xi|s)$ but does have an impact on $\overline{\overline{H}}(\Xi)$.

If, on the other hand, $p_2 \geq p_1$, then $\Delta C = 0$ regardless whether the uncertainty set is bounded or not, so that, in general,

$$\Delta C = C_w - C_{FB} = [h(p_1) - h(p_2)]_+ \quad (66)$$

where $[x]_+ = \max\{0, x\}$.

C. Example 3

Let the compound noise sequence be of the form

$$\xi_s^n = \{w_1, \dots, w_n\} \quad (67)$$

with binary alphabet and $W_i \sim \text{Ber}(p_i)$ and independent of each other, where

$$p_i = \frac{s}{2(i+s)} \quad (68)$$

and $s \geq 0$ (not necessarily integer). This models a scenario where noise becomes "weaker" with time (note that $h(p_i)$ decreases with i), while s controls the decay rate: noise becomes negligible when $i \gg s$, so that $h(p_i) \approx 0$, see Fig. 2. The process is clearly not stationary.

After some manipulations, one obtains $\overline{H}(\Xi|s) = 0 \ \forall \ s$ and hence $\sup_s \overline{H}(\Xi|s) = 0$. Yet, $\overline{\overline{H}}(\Xi) = 1$ so that $C_w = 1$, $C_{FB} = 0$ and $\Delta C = 1$, the maximal possible value, and the noise is not uniform. Thus, while the Tx CSI is the most useful, the noiseless causal feedback is useless.

As above, bounding the uncertainty set $s \leq S < \infty$ changes the situation dramatically: $C_{FB} = 1$, $\Delta C = 0$, so that the Tx CSI gives no increase in the capacity since the noise is now uniform.

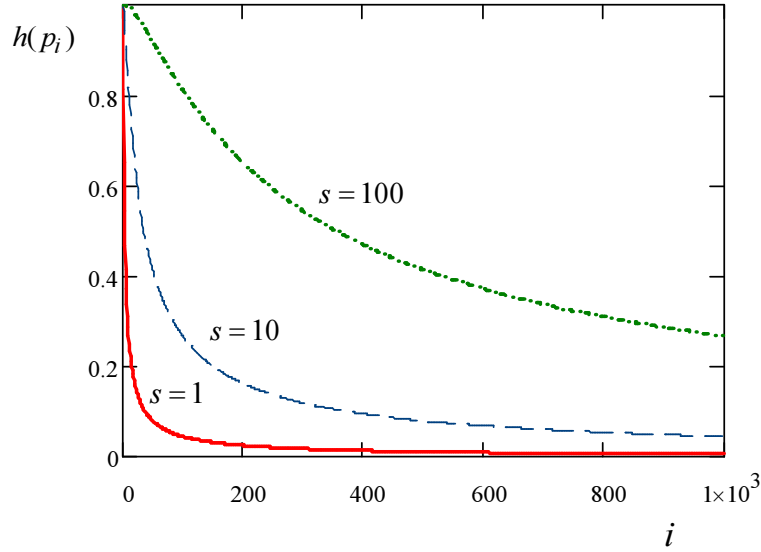


Fig. 2. Binary entropy $h(p_i)$ for p_i as in (68) versus time i for different states s .

D. Example 4

Let us now consider a non-ergodic non-stationary channel with

$$\xi_s^n = \begin{cases} w_s^n & \text{with } \Pr = p \\ z_s^n & \text{with } \Pr = 1 - p \end{cases} \quad (69)$$

where W_s^n, Z_s^n are non-stationary processes (e.g. from the above examples) and $0 < p < 1$, i.e. one of the two processes is randomly selected at the beginning and it operates during the entire transmission. This process is clearly non-ergodic when its components are of different distributions. One may also consider p_s , i.e. a function of the channel state, provided that $0 < \alpha \leq p_s \leq \beta < 1 \forall s$.

It can be seen that

$$\overline{H}(\Xi|s) = \max\{\overline{H}(\mathbf{W}|s), \overline{H}(\mathbf{Z}|s)\}, \quad \overline{\overline{H}}(\Xi) = \max\{\overline{\overline{H}}(\mathbf{W}), \overline{\overline{H}}(\mathbf{Z})\} \quad (70)$$

so that

$$\Delta C = \max\{\overline{\overline{H}}(\mathbf{W}), \overline{\overline{H}}(\mathbf{Z})\} - \sup_s \max\{\overline{H}(\mathbf{W}|s), \overline{H}(\mathbf{Z}|s)\} \quad (71)$$

In particular, $\Delta C = 0$ if $\{W_s^n\}_{n=1}^\infty, \{Z_s^n\}_{n=1}^\infty$ are uniform compound sequences. This holds if e.g. the uncertainty set is of a finite cardinality, regardless of what the distributions of $\{W_s^n\}, \{Z_s^n\}$ are.

VII. STRONG CONVERSE

In this section, we establish a sufficient and necessary condition for the strong converse to hold for the compound channel with additive noise. In addition to being of theoretical interest on its own, it also has some practical implications. In particular, strong converse ensures that slightly larger error probability cannot be traded off for higher data rate, since the transition from arbitrary low to high error probability is sharp. Additionally, a consequence of this is that the error rate performance degrades dramatically if the SNR drops below the threshold for which the system was designed.

Let ε_n and r_n be the error probability and rate of a codebook of blocklength n . The formal definition of strong converse is as follows.

Definition 4. A compound channel is said to satisfy strong converse if

$$\lim_{n \rightarrow \infty} \varepsilon_n = 1 \quad (72)$$

for any code satisfying

$$\liminf_{n \rightarrow \infty} r_n > C_c \quad (73)$$

We begin with the following definitions which are needed below. 1st one extends the standard definition of convergence in probability to compound random sequences.

Definition 5. A compound random sequence $\{Y_{sn}\}_{n=1}^{\infty}$ is said to converge in probability to y_0 , denoted as $Y_{sn} \xrightarrow{\Pr} y_0$, if

$$\lim_{n \rightarrow \infty} \sup_s \Pr\{|Y_{sn} - y_0| > \epsilon\} = 0 \quad (74)$$

for any $\epsilon > 0$, where \sup_s is over the whole state set.

It should be emphasized that the point-wise convergence, i.e. $\lim_{n \rightarrow \infty} \Pr\{|Y_{sn} - y_0| > \epsilon\} = 0 \forall s$, does not imply (74), which is a stronger condition (see also (55)).

In addition to the following standard definitions of the infimum \underline{X}_s and supremum \overline{X}_s of a random sequence X_s^n under state s [2][3]:

$$\begin{aligned} \underline{X}_s &= \sup \left\{ x : \lim_{n \rightarrow \infty} \Pr\{X_{sn} \leq x\} = 0 \right\} \\ \overline{X}_s &= \inf \left\{ x : \lim_{n \rightarrow \infty} \Pr\{X_{sn} \geq x\} = 0 \right\} \end{aligned} \quad (75)$$

and the compound infimum $\underline{\underline{X}}$ and supremum $\overline{\overline{X}}$ in Definition 1, the following compound inf and sup operators are needed in a condition for strong converse.

Definition 6. Let $\{X_{sn}\}_{n=1}^{\infty}$ be an arbitrary compound random sequence where s is a state. The compound infimum $\underline{\cdot}$ and supremum $\overline{\cdot}$ operators are defined as follows:

$$\underline{\underline{X}} = \underline{\underline{X_{sn}}} = \sup \left\{ x : \lim_{n \rightarrow \infty} \inf_s \Pr \{X_{sn} \leq x\} = 0 \right\} \quad (76)$$

$$\overline{\overline{X}} = \overline{\overline{X_{sn}}} = \inf \left\{ x : \lim_{n \rightarrow \infty} \inf_s \Pr \{X_{sn} \geq x\} = 0 \right\} \quad (77)$$

Roughly, $\underline{\underline{X}}_s$ and $\overline{\overline{X}}_s$ represent the largest lower and least upper bounds of the asymptotic support set of X_{sn} under state s while $\underline{\underline{X}}$ and $\overline{\overline{X}}$ do so over the whole state set by selecting the best states for the respective bounds. Note however that these quantities are different from $\underline{\underline{X}}$ and $\overline{\overline{X}}$: inf rather than sup are used in the definitions of $\underline{\underline{X}}$ and $\overline{\overline{X}}$ so that the respective limits are enforced for some channel states only, not over the whole state set. While subtle, the difference is important, as we will see below. These operators have the properties which are instrumental in establishing the strong converse and other results.

Proposition 2. The compound inf and sup operators in Definition 6 satisfy the following:

$$\underline{\underline{(-X)}} = -\overline{\overline{X}} \quad (78)$$

$$\overline{\overline{X}} + \underline{\underline{Y}} \leq \overline{\overline{(X + Y)}} \leq \overline{\overline{X}} + \overline{\overline{Y}} \quad (79)$$

$$\underline{\underline{X}} \leq \min\{\underline{\underline{X}}, \overline{\overline{X}}\} \leq \max\{\underline{\underline{X}}, \overline{\overline{X}}\} \leq \overline{\overline{X}} \quad (80)$$

$$\sup_s \underline{\underline{X}}_s \leq \underline{\underline{X}}, \quad \overline{\overline{X}} \leq \inf_s \overline{\overline{X}}_s \quad (81)$$

If $Y_{sn} \xrightarrow{\Pr} y_0$, then

$$\overline{\overline{(X + Y)}} = \overline{\overline{X}} + y_0 \quad (82)$$

Proof: See Appendix. ■

Strict inequalities in Proposition 2 can be demonstrated via examples.

Example 1: Let X_{1n} and X_{2n} be uniformly-distributed random variables,

$$X_{1n} \sim \text{uni}[0, 2], \quad X_{2n} \sim \text{uni}[1, 3] \quad (83)$$

so that

$$\underline{\underline{X}} = 0, \quad \underline{\underline{X}} = 1, \quad \overline{\overline{X}} = 2, \quad \overline{\overline{X}} = 3 \quad (84)$$

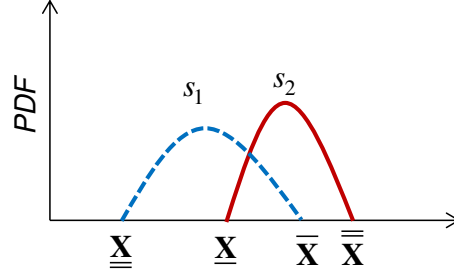


Fig. 3. Asymptotic distribution of a random 2-state sequence \underline{X} and related quantities. Note that $\underline{X} < \overline{X}$.

and all inequalities in (80) are strict. Since

$$\sup_s \underline{X}_s = 1, \quad \inf_s \overline{X}_s = 2, \quad (85)$$

this example also demonstrates that the inequalities in (81) can become equalities.

To demonstrate that the inequalities in (79) can be strict, set X_{sn} to be deterministic constant and Y_{sn} as X in Example 1 above.

Example 2: To see that the inequalities in (81) can be strict, let X_{sn} be Bernoulli random variables as follows:

$$X_{sn} \sim \text{Ber}\{1 - p_{sn}\}, \quad p_{sn} = \frac{n}{n+s}, \quad s \geq 0 \quad (86)$$

so that $\Pr\{X_{sn} = 0\} = p_{sn}$. It is straightforward to see that $\underline{X}_s = \overline{X}_s = 0$ for any s so that

$$\sup_s \underline{X}_s = \inf_s \overline{X}_s = 0, \quad (87)$$

yet $\underline{X} = 1$ so that 1st inequality is strict while 2nd one becomes equality since $\overline{X} = 0$. To see that this inequality can be strict, set $X_{sn} \sim \text{Ber}\{p_{sn}\}$ instead, so that

$$\sup_s \underline{X}_s = \inf_s \overline{X}_s = 1, \quad (88)$$

yet $\overline{X} = 0$.

Using Example 1 and its modifications, see Fig. 3 and 4, one can also demonstrate that there is no specific relationship between \underline{X} and \overline{X} in general, i.e. neither $\underline{X} \leq \overline{X}$ nor $\underline{X} \geq \overline{X}$ are true, unlike $\underline{\underline{X}} \leq \overline{\overline{X}}$ that holds in full generality. In a similar way, it can be shown that there exists no specific relationship between $\sup_s \overline{X}_s$ and \underline{X} . This also holds for $\inf_s \underline{X}_s$ and \overline{X} .

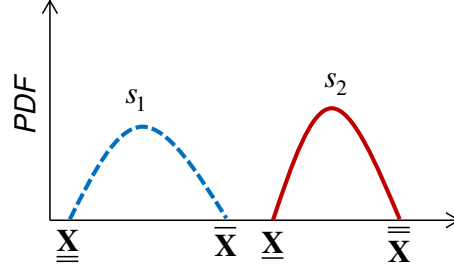


Fig. 4. Asymptotic distribution of a random 2-state sequence \underline{X} and related quantities. Note that $\underline{X} > \overline{X}$.

Using Proposition 9 in [8] and (78), the inequalities in (81) can be refined as follows:

$$\begin{aligned} \underline{\underline{X}} &\leq \inf_s \underline{X}_s \leq \sup_s \underline{X}_s \leq \underline{X} \leq \overline{\overline{X}}, \\ \underline{\underline{X}} &\leq \overline{X} \leq \inf_s \overline{X}_s \leq \sup_s \overline{X}_s \leq \overline{\overline{X}} \end{aligned} \quad (89)$$

A special case of (79) is when $Y_{sn} = b$, i.e. a constant, so that, for any $a \geq 0$,

$$\overline{(a\underline{X} + b)} = a\overline{X} + b \quad (90)$$

i.e. $\overline{\{\cdot\}}$ is a linear operator for positive a . It is straightforward to see that, for negative a ,

$$\overline{(a\underline{X} + b)} = a\underline{X} + b \quad (91)$$

Let $\underline{H}(\Xi) = \{n^{-1}h(\Xi_s^n|s)\}$ and likewise for $\overline{H}(\Xi)$. In addition to its properties inherited from Proposition 2, it also satisfies

$$0 \leq \underline{H}(\Xi), \overline{H}(\Xi) \leq \log M \quad (92)$$

where 1st inequality holds in full generality and 2nd one - for M -ary alphabets. We are now in a position to establish a sufficient and necessary condition for the strong converse to hold.

Theorem 4. *The compound channel with additive noise in (1) under the full Rx CSI satisfies the strong converse condition for both feedback and no feedback cases if and only if*

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) \quad (93)$$

If the compound noise is uniform, this reduces to

$$\sup_s \overline{H}(\Xi|s) = \underline{H}(\Xi) \quad (94)$$

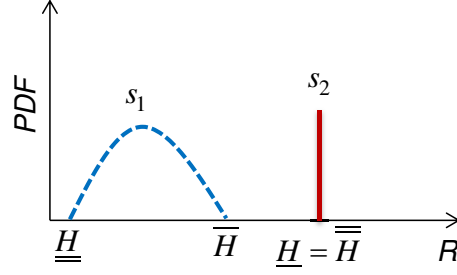


Fig. 5. Asymptotic distribution of the noise entropy density rate for a 2-state channel with strong converse and related entropy density rates.

Proof: See Appendix. ■

Fig. 5 illustrates the condition of strong converse for a 2-state channel.

Using Proposition 27 in [8] under the optimal (uniform) input \mathbf{X}^* in combination with (130), one further obtains under the strong converse condition (93):

$$\overline{\overline{H}}(\Xi) = \limsup_{n \rightarrow \infty} \sup_s \frac{1}{n} H(\Xi_s^n) \quad (95)$$

where $H(\Xi_s^n)$ is the ergodic entropy, i.e. the compound sup-entropy rate $\overline{\overline{H}}(\Xi)$ coincides with the ergodic entropy rate of the noise (under its worst states), even though no ergodicity (or information stability) was imposed on the noise upfront⁴. Hence, one concludes that the strong converse condition forces the worst-case noise to behave ergodically and hence the worst-case noise ergodicity is both necessary and sufficient for the strong converse to hold. This conclusion holds for both feedback and no feedback cases.

While there is no specific ordering between $\underline{H}(\Xi)$ and $\overline{H}(\Xi)$ or between $\sup_s \overline{H}(\Xi|s)$ and $\underline{H}(\Xi)$ in general (as indicated by the examples above), such ordering is induced by the strong converse, as indicated below.

Corollary 4.1. *Under the strong converse condition in Theorem 4, the following ordering holds:*

$$\overline{H}(\Xi) \leq \inf_s \overline{H}(\Xi|s) \leq \sup_s \overline{H}(\Xi|s) \leq \underline{H}(\Xi) \quad (96)$$

which is thus a necessary condition for the strong converse to hold.

⁴Note also that (95) equates two very different quantities: while the definition of $H(\Xi_s^n)$ is based on the expectation, so it is an ergodic quantity, that of $\overline{\overline{H}}(\Xi)$ does not use expectation at all.

Proof: It follows from (93) and (81) that

$$\begin{aligned}
 \underline{H}(\Xi) = \overline{\overline{H}}(\Xi) &\geq \sup_s \overline{H}(\Xi|s) \\
 &\geq \inf_s \overline{H}(\Xi|s) \\
 &\geq \overline{H}(\Xi)
 \end{aligned} \tag{97}$$

■

A. Examples

To gain further insight, one may use the examples of Section VI. In particular, one obtains for Example 1

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = \log M \tag{98}$$

when the uncertainty set is not bounded and

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = 0 \tag{99}$$

when it is, so that the strong converse holds in both cases.

For Example 2,

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = h(p_1) \tag{100}$$

when the uncertainty set is not bounded and

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = h(p_2) \tag{101}$$

when it is, so that the strong converse holds in both cases as well.

For Example 3,

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = 1 \tag{102}$$

when the uncertainty set is not bounded and

$$\overline{\overline{H}}(\Xi) = \underline{H}(\Xi) = 0 \tag{103}$$

when it is, so that the strong converse holds in both cases too.

Example 4 is more interesting. It is not too difficult to show that, in the general case,

$$\underline{H}(\Xi) \leq \min\{\underline{H}(\mathbf{W}), \underline{H}(\mathbf{Z})\} \tag{104}$$

so that

$$\begin{aligned}
\underline{H}(\Xi) &\leq \min\{\underline{H}(\mathbf{W}), \underline{H}(\mathbf{Z})\} \\
&\leq \max\{\underline{H}(\mathbf{W}), \underline{H}(\mathbf{Z})\} \\
&\leq \overline{\overline{H}}(\Xi) = \max\{\overline{\overline{H}}(\mathbf{W}), \overline{\overline{H}}(\mathbf{Z})\}
\end{aligned} \tag{105}$$

and hence, if $\underline{H}(\mathbf{W}) \neq \underline{H}(\mathbf{Z})$,

$$\underline{H}(\Xi) < \overline{\overline{H}}(\Xi) \tag{106}$$

so that the strong converse does not hold (one may use Examples 1-3 to construct component sequences \mathbf{W}, \mathbf{Z} for further insights). Note that this conclusion holds for any p as long as $0 < p < 1$.

Remark 5. *It is tempting to conclude, based on $\underline{\underline{H}}(\Xi) = \min\{\underline{\underline{H}}(\mathbf{W}), \underline{\underline{H}}(\mathbf{Z})\}$ which holds in full generality, that (104) should hold with equality in general. To see that this is not the case, consider Example 4 with the following component sequences:*

$$\begin{aligned}
w_s^n &= \{b_1..b_s, 0..0\} \\
z_s^n &= \{0..0, b_{s+1}..b_n\}
\end{aligned} \tag{107}$$

where b^n is a binary i.i.d. equiprobable sequence. This models a scenario where the noise randomly corrupts either 1st or 2nd part of a codeword and s controls its length. It follows that

$$\overline{\overline{H}}(\Xi) = \underline{H}(\mathbf{W}) = \underline{H}(\mathbf{Z}) = 1 \tag{108}$$

yet

$$\underline{H}(\Xi) = 1/2 < 1 = \min\{\underline{H}(\mathbf{W}), \underline{H}(\mathbf{Z})\} \tag{109}$$

Note that the strong converse does not hold in this case either, even though it holds for each component sequence individually and $\underline{H}(\mathbf{W}) = \underline{H}(\mathbf{Z})$. Further note that $\overline{\overline{H}}(\Xi) = 1/2$, $\inf_s \overline{\overline{H}}(\Xi|s) = \sup_s \overline{\overline{H}}(\Xi|s) = 1$ so that the last inequality in (96) does not hold.

VIII. CONCLUSION

The capacity of compound channels with additive noise and the Rx CSI has been studied. When all alphabets are discrete and there is no cost constraint, noiseless causal feedback does not increase the capacity. The impact of the channel state information at the transmitter has been quantified. In particular, it does not increase the capacity if the additive noise is a uniform compound process. Otherwise, it may provide significant improvement (unlike the feedback), which was shown via examples. A saddle-point has been shown to exist in the information-theoretic game between the transmitter and the nature, even though the objective is not convex/concave in the right way. Finally, the sufficient and necessary condition for the strong converse to hold has been established: it requires the worst-case noise sequence to behave ergodically, even though no ergodicity or information satiability requirements were imposed upfront. Examples are provided to facilitate understanding and insights.

IX. ACKNOWLEDGEMENT

The authors are grateful to A. Lapidoth and E. Telatar for their support, and to P. Mitran and M. Raginsky for fruitful discussions.

X. APPENDIX

A. Proof of Lemma 2

The proof of the 1st part (the inequality in general) is by contradiction. Assume that $\overline{\overline{H}}(\Xi) < \sup_s \overline{H}(\Xi|s)$, which implies that

$$\exists s_0 : \overline{\overline{H}} = \overline{\overline{H}}(\Xi) < \overline{H} = \overline{H}(\Xi|s_0) \quad (110)$$

Set

$$R = (\overline{\overline{H}} + \overline{H})/2 = \overline{\overline{H}} + \Delta = \overline{H} - \Delta \quad (111)$$

where $\Delta = (\overline{H} - \overline{\overline{H}})/2 > 0$. Note that

$$\lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} h(\Xi_{s_0}^n | s_0) > \overline{H} - \Delta \right\} > 0 \quad (112)$$

from the definition of \overline{H} . However,

$$\begin{aligned}
0 &= \lim_{n \rightarrow \infty} \sup_s \Pr \left\{ \frac{1}{n} h(\Xi_s^n | s) > \overline{H} + \Delta \right\} \\
&\geq \lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} h(\Xi_{s_0}^n | s_0) > \overline{H} + \Delta \right\} \\
&= \lim_{n \rightarrow \infty} \Pr \left\{ \frac{1}{n} h(\Xi_{s_0}^n | s_0) > \overline{H} - \Delta \right\} > 0
\end{aligned} \tag{113}$$

where 1st equality is due to the definition of \overline{H} , i.e. a contradiction, from which the desired inequality follows.

The "if" part of the equality case (under uniform noise) is also proved by contradiction: assume that, under the uniform convergence,

$$\overline{\overline{H}} > \overline{H} = \sup_s \overline{H}(\Xi | s) \tag{114}$$

and set

$$R = (\overline{\overline{H}} + \overline{H})/2 = \overline{\overline{H}} - \Delta = \overline{H} + \Delta \tag{115}$$

where $\Delta = (\overline{\overline{H}} - \overline{H})/2 > 0$, and hence

$$\lim_{n \rightarrow \infty} \Pr \{ n^{-1} h(\Xi_s^n | s) > \overline{H} + \Delta \} = 0 \quad \forall s \in \mathcal{S} \tag{116}$$

from the definition of \overline{H} , so that a contradiction follows

$$\begin{aligned}
0 &= \sup_s \lim_{n \rightarrow \infty} \Pr \{ n^{-1} h(\Xi_s^n | s) > \overline{H} + \Delta \} \\
&= \lim_{n \rightarrow \infty} \sup_s \Pr \{ n^{-1} h(\Xi_s^n | s) > \overline{H} + \Delta \} \\
&= \lim_{n \rightarrow \infty} \sup_s \Pr \left\{ n^{-1} h(\Xi_s^n | s) > \overline{\overline{H}} - \Delta \right\} > 0
\end{aligned} \tag{117}$$

where 2nd equality is due to uniform convergence and the last inequality is from the definition of $\overline{\overline{H}}$.

To prove the "only if" part, assume that the equality holds and observe that

$$\begin{aligned}
0 &= \lim_{n \rightarrow \infty} \sup_s \Pr \left\{ n^{-1} h(\Xi_s^n | s) > \overline{\overline{H}} + \Delta \right\} \\
&= \lim_{n \rightarrow \infty} \sup_s \Pr \left\{ n^{-1} h(\Xi_s^n | s) > \sup_s \overline{H}(\Xi | s) + \Delta \right\}
\end{aligned} \tag{118}$$

for any $\Delta > 0$. The last equality implies uniform convergence: for any $\epsilon > 0$ there exists such $n_0(\epsilon)$ that for any $n > n_0(\epsilon)$,

$$\sup_s \Pr \left\{ n^{-1} h(\Xi_s^n | s) > \sup_s \overline{H}(\Xi | s) + \Delta \right\} < \epsilon$$

and hence the convergence is uniform.

B. Proof of Proposition 2

Let $\liminf = \lim_{n \rightarrow \infty} \inf_s$ and likewise for \limsup . Eq. (78) follows from the definition of $\underline{\{\cdot\}}$:

$$\begin{aligned} \underline{(-\mathbf{X})} &= \sup \{x : \liminf \Pr \{-X_{sn} \leq x\} = 0\} \\ &= \sup \{x : \liminf \Pr \{X_{sn} \geq -x\} = 0\} \\ &= \sup \{-z : \liminf \Pr \{X_{sn} \geq z\} = 0\} \\ &= -\inf \{z : \liminf \Pr \{X_{sn} \geq z\} = 0\} \\ &= -\overline{\mathbf{X}} \end{aligned} \tag{119}$$

To prove (79), set $x = \overline{\mathbf{X}} + \overline{\overline{\mathbf{Y}}} + \delta$ for some $\delta > 0$, let B denote the event $\{Y_{sn} < \overline{\overline{\mathbf{Y}}} + \delta\}$ and B^c - its complement, and observe that

$$\begin{aligned} 0 &= \liminf \Pr \{X_{sn} + \overline{\overline{\mathbf{Y}}} \geq x\} \\ &= \liminf (\Pr \{X_{sn} + \overline{\overline{\mathbf{Y}}} \geq x | B\} \Pr \{B\} + \Pr \{X_{sn} + \overline{\overline{\mathbf{Y}}} \geq x | B^c\} \Pr \{B^c\}) \\ &\geq \liminf \Pr \{X_{sn} + \overline{\overline{\mathbf{Y}}} \geq x | B\} \Pr \{B\} \\ &\geq \liminf \Pr \{X_{sn} + Y_{sn} - \delta \geq x | B\} \Pr \{B\} \\ &= \liminf \Pr \{X_{sn} + Y_{sn} - \delta \geq x | B\} \Pr \{B\} + \limsup \Pr \{X_{sn} + Y_{sn} - \delta \geq x | B^c\} \Pr \{B^c\} \\ &\geq \liminf (\Pr \{X_{sn} + Y_{sn} - \delta \geq x | B\} \Pr \{B\} + \Pr \{X_{sn} + Y_{sn} - \delta \geq x | B^c\} \Pr \{B^c\}) \\ &= \liminf \Pr \{X_{sn} + Y_{sn} \geq x + \delta\} = 0 \end{aligned} \tag{120}$$

where 1st equality is from $x = \overline{\mathbf{X}} + \overline{\overline{\mathbf{Y}}} + \delta$ and the definition of $\overline{\mathbf{X}}$; 2nd inequality is from $\overline{\overline{\mathbf{Y}}} > Y_{sn} - \delta$ conditioned on B ; 3rd equality is from

$$\limsup \Pr \{X_{sn} + Y_{sn} - \delta \geq x | B^c\} \Pr \{B^c\} \leq \limsup \Pr \{B^c\} = 0 \tag{121}$$

where the last equality is from the definition of B^c ; the last equality in (120) is implied by the preceding chain. This last equality implies that $\overline{\mathbf{X} + \mathbf{Y}} \leq x + \delta$ so that

$$\overline{\mathbf{X} + \mathbf{Y}} \leq \overline{\mathbf{X}} + \overline{\mathbf{Y}} + 2\delta \quad (122)$$

for any $\delta > 0$, which proves 2nd inequality in (79). To prove 1st one, use the substitutions $\mathbf{Y} \rightarrow -\mathbf{Y}$ and $\mathbf{X} \rightarrow \mathbf{X} + \mathbf{Y}$ in combination with (78).

To establish (80), we first show that $\underline{\underline{\mathbf{X}}} \leq \underline{\mathbf{X}}$. To this end, let

$$\begin{aligned} \Omega_1 &= \{x : \limsup \Pr \{X_{sn} \leq x\} = 0\} \\ \Omega_2 &= \{x : \liminf \Pr \{X_{sn} \leq x\} = 0\} \end{aligned} \quad (123)$$

Since

$$\limsup \Pr \{X_{sn} \leq x\} \geq \liminf \Pr \{X_{sn} \leq x\} \quad (124)$$

it follows that $\Omega_1 \subset \Omega_2$, which implies $\underline{\underline{\mathbf{X}}} \leq \underline{\mathbf{X}}$ by using sup. Next, we show that $\underline{\underline{\mathbf{X}}} \leq \overline{\mathbf{X}}$. To this end, let

$$\Omega_3 = \{x : \limsup \Pr \{X_{sn} \leq x\} = 1\}$$

and observe that

$$\begin{aligned} \overline{\mathbf{X}} &= \inf \{x : \liminf \Pr \{X_{sn} \geq x\} = 0\} \\ &= \inf \{x : \liminf \Pr \{X_{sn} > x\} = 0\} \\ &= \inf \{x \in \Omega_3\} \end{aligned} \quad (125)$$

Since, for any $x_1 \in \Omega_1$ and any $x_3 \in \Omega_3$, it holds that $x_1 < x_3$, so that

$$\underline{\underline{\mathbf{X}}} = \sup \{x \in \Omega_1\} \leq \inf \{x \in \Omega_3\} = \overline{\mathbf{X}} \quad (126)$$

This establishes 1st inequality in (80). 2nd one is trivial. 3rd one can be established from 1st one using $\mathbf{X} \rightarrow -\mathbf{X}$.

To show 1st inequality in (81), recall that

$$\underline{\underline{\mathbf{X}}}_s = \sup \left\{ x : \lim_{n \rightarrow \infty} \Pr \{X_{sn} \leq x\} = 0 \right\}, \quad (127)$$

set $x_0 = \underline{\underline{\mathbf{X}}}_s - \delta$ for some $\delta > 0$ and observe that

$$0 = \lim_{n \rightarrow \infty} \Pr \{X_{sn} \leq x_0\} \geq \liminf \Pr \{X_{sn} \leq x_0\} = 0 \quad (128)$$

where the last equality is implied by the preceding chain. This implies that $\underline{\mathbf{X}} \geq x_0$. Since this holds for any $\delta > 0$, $\underline{\mathbf{X}} \geq \underline{\mathbf{X}}_s$ follows. Since this holds for any s , 1st inequality in (81) follows. 2nd one can be established via $\mathbf{X} \rightarrow -\mathbf{X}$.

To establish (82), observe that $Y_{sn} \xrightarrow{\text{Pr}} y_0$ implies $\underline{\underline{\mathbf{Y}}} = \overline{\overline{\mathbf{Y}}} = y_0$ and use (79).

C. Proof of Theorem 4

We begin with a brief summary of the sufficient and necessary condition for the general compound channel to satisfy the strong converse.

Theorem 5 ([8][9]). *The general compound channel with full Rx CSI and without feedback satisfies the strong converse condition if and only if*

$$C_c \triangleq \sup_{\mathbf{X}} \underline{\underline{I}}(\mathbf{X}; \mathbf{Y}) = \sup_{\mathbf{X}} \overline{\overline{I}}(\mathbf{X}; \mathbf{Y}) \quad (129)$$

where \sup is over all sequences of finite-dimensional input distributions. The condition (129) is equivalent to the following: for any $\delta > 0$ and an optimal input \mathbf{X}^* ,

$$\lim_{n \rightarrow \infty} \inf_s \Pr\{|Z_{ns}^* - C_c| > \delta\} = 0 \quad (130)$$

where $Z_{ns}^* = \frac{1}{n} i(X^{n*}; Y^{n*} | s)$ is the information density rate under optimal input \mathbf{X}^* , i.e. there exists such sequence of channel states $s(n)$ that the corresponding information density rate Z_{ns}^* under optimal input \mathbf{X}^* converges in probability to the compound capacity C_c (i.e. the channel represented by this sequence of states is information-stable, even though the original compound channel is not required to be information-stable).

To adapt this result to the feedback case, we again consider \mathbf{W} as an input and optimize over both \mathbf{W} and \mathbf{F} so that (129) becomes

$$\sup_{\mathbf{W}, \mathbf{F}} \underline{\underline{I}}(\mathbf{W}; \mathbf{Y}) = \sup_{\mathbf{W}, \mathbf{F}} \overline{\overline{I}}(\mathbf{W}; \mathbf{Y}) \quad (131)$$

Since the left-hand side has been already evaluated, we now evaluate the right-hand side. To this end, one can follow the steps similar to those in evaluating the left-hand side. First, observe that

$$\begin{aligned} \sup_{\mathbf{W}, \mathbf{F}} \overline{\overline{I}}(\mathbf{W}; \mathbf{Y}) &\leq \sup_{\mathbf{W}, \mathbf{F}} [\overline{\overline{H}}(\mathbf{Y}) - \underline{\underline{H}}(\mathbf{Y} | \mathbf{W})] \\ &\leq \log M - \inf_{\mathbf{W}, \mathbf{F}} \underline{\underline{H}}(\mathbf{Y} | \mathbf{W}) \\ &= \log M - \underline{\underline{H}}(\mathbf{\Xi}) \end{aligned} \quad (132)$$

where 1st inequality is due to Proposition 2; 2nd inequality follows from $\overline{\overline{H}}(\mathbf{Y}) \leq \log M$ (since the alphabet is M -ary); the last equality is due to (42) so that $\underline{H}(\mathbf{Y}|\mathbf{W}) = \underline{H}(\Xi)$. Now, using no feedback and uniform input \mathbf{X} , one obtains $\overline{I}(\mathbf{W}; \mathbf{Y}) = \log M - \underline{H}(\Xi)$ so that

$$\sup_{\mathbf{W}, \mathbf{F}} \overline{I}(\mathbf{W}; \mathbf{Y}) \geq \log M - \underline{H}(\Xi) \quad (133)$$

Combining the two inequalities,

$$\sup_{\mathbf{W}, \mathbf{F}} \overline{I}(\mathbf{W}; \mathbf{Y}) = \log M - \underline{H}(\Xi) \quad (134)$$

It is remarkable that, similarly to $\underline{I}(\mathbf{W}; \mathbf{Y})$, the optimal value of $\overline{I}(\mathbf{W}; \mathbf{Y})$ is not affected by feedback either and the best strategy is to use the uniformly-distributed input and ignore feedback. Combining the last equality with (26), the desired condition follows.

REFERENCES

- [1] E. Biglieri, J. Proakis, and S. Shamai, "Fading Channels: Information-Theoretic and Communications Aspects," *IEEE Trans. Inform. Theory*, vol. 44, No. 6, pp. 2619-2692, Oct. 1998.
- [2] S. Verdú, T.S. Han, "A General Formula for Channel Capacity", *IEEE Trans. Info. Theory*, vol. 40, no. 4, pp. 1147-1157, July 1994.
- [3] T. S. Han, *Information-Spectrum Method in Information Theory*, New York: Springer, 2003.
- [4] A. Lapidoth and P. Narayan, "Reliable Communication Under Channel Uncertainty," *IEEE Trans. Info. Theory*, vol. 44, No. 6, Oct. 1998.
- [5] B. Shrader, H. Permuter, Feedback Capacity of the Compound Channel, *IEEE Trans. Info. Theory*, v. 55, no. 8, pp. 3629-3644, Aug. 2009.
- [6] S. Loyka, C. D. Charalambous, A General Formula for Compound Channel Capacity, *IEEE Int. Symp. on Information Theory (ISIT-15)*, Hong Kong, June 14-19, 2015.
- [7] F. Alajaji, Feedback Does Not Increase the Capacity of Discrete Channels with Additive Noise, *IEEE Trans. Info. Theory*, v. 41, N. 2, pp. 546-549, Mar. 1995.
- [8] S. Loyka, C. D. Charalambous, A General Formula for Compound Channel Capacity, *IEEE Trans. Info. Theory*, v. 62, n. 7, pp. 3971-3991, July 2016.
- [9] S. Loyka, C. D. Charalambous, Strong Converse for General Compound Channels, *International Zurich Seminar on Communications (IZS)*, March 2 - 4, 2016, Zurich, Switzerland.
- [10] G. Kramer, Capacity Results for the Discrete Memoryless Network, *IEEE Trans. Info. Theory*, v. 49, no. 1, pp 4-21, Jan. 2003.
- [11] S. Tatikonda, S. Mitter, The Capacity of Channels with Feedback, *IEEE Trans. Info. Theory*, v.55, no. 1, pp. 323-349, Jan. 2009.
- [12] R. L. Dobrushin, "A General Formulation of The Fundamental Theorem of Shannon in Information Theory", *Uspekhi Mat. Nauk*, v. 14, no. 6(90), Nov.-Dec. 1959, pp.3-104.

- [13] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. San Francisco: Holden-Day, 1964.
- [14] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 2006.
- [15] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [16] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, vol. I: Fixed-Point Theorems, Springer, New York, 1986.
- [17] R.L. Dobrushin, "Optimal Information Transmission Through a Channel With Unknown Parameters," *Radiotekhnika i Elektronika*, vol. 4, n. 12, pp. 1951–1956, 1959.
- [18] R.J. McEliece, *Communications in The Presence of Jamming: An Information-Theoretic Approach*, in G. Longo (Ed.), *Secure Digital Communications*, Springer, Wien, 1983.
- [19] J.M. Borden, D.M. Mason, R.J. McEliece, Some Information Theoretic Saddlepoints, *SIAM J. Control Optim.*, v. 23, no. 1, pp. 129-143, Jan. 1985.
- [20] S. Diggavi, T.M. Cover, The Worst Additive Noise Under a Covariance Constraint, *IEEE Trans. Info. Theory*, v. 47, n. 7, pp. 3072–3081, Nov. 2001.
- [21] R. Mathar, A. Schmeink, Saddle Point Properties and Nash Equilibria for Channel Games, *EURASIP Journal on Advances in Signal Processing*, vol. 2009, article ID 823513, pp. 1-9, Jan. 2009.
- [22] S. Loyka, C. D. Charalambous, Novel Matrix Singular Value Inequalities and Their Applications to Uncertain MIMO Channels, *IEEE Trans. Info. Theory*, v. 61, N. 12, pp. 6623 - 6634, Dec. 2015.